Information Sciences 273 (2014) 156–170

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Fuzzy control GA with a novel hybrid semantic similarity strategy for text clustering

Wei Song [a,b,*], Jiu Zhen Liang [a], Soon Cheol Park [b]

[a] School of Internet of Things Engineering, Jiangnan University, Wuxi, Jiangsu 214122, The People's Republic of China
[b] School of Information & Communication Engineering, Chonbuk National University, Jeonju, Jeonbuk 561-756, Republic of Korea

ARTICLE INFO

Article history:
Received 20 September 2008
Received in revised form 5 December 2011
Accepted 8 March 2014
Available online 18 March 2014

Keywords:
Clustering
WordNet
Hybrid semantic similarity
Fuzzy control
Evolutionary computation
Genetic algorithm

ABSTRACT

This paper proposes a fuzzy control genetic algorithm (GA) in conjunction with a novel hybrid semantic similarity measure for document clustering. Since the common clustering algorithms use vector space model (VSM) to represent document, the conceptual relationships between related terms being ignored, we use semantic similarity measures to solve this problem. In general, the semantic similarity measures can be extensively categorized into two kinds: thesaurus-based methods and corpus-based methods. However, in practice the corpus-based method is rather complicated to tackle. We propose and demonstrate a semantic space model (SSM) as the corpus-based method, where the appropriately reduced dimensions in SSM can capture the true relationship between documents in terms of concepts, rather than specific terms. Thus, the thesaurus-based method is combined with our SSM as a hybrid strategy to represent the semantic similarity measure. In GA field, the balance between the capability to converge to an optimum and the capacity to explore new solutions affects the success of search for the global optimum. We utilize a fuzzy control GA to adaptively adjust the influence between these two factors. Two textual data sets from Reuter document collection and 20-newsgroup corpus are tested in our experiments, and the results show that our fuzzy control GA combined with the hybrid semantic similarity strategy apparently outperforms the conventional GA, FCM and K-means with the traditional cosine similarity in VSM. Moreover, the superiorities of the fuzzy control GA and our hybrid semantic strategy are demonstrated by their better performance, in comparison with conventional GA with the same similarity measures.

1. Introduction

Owing to the overflowing of knowledge and information on the internet, it becomes impossible to manually cluster documents online. Thus automatic and high-quality partitioning data into previously unseen categories turns to be a major topic for applications such as data mining and information retrieval. Clustering [13,26,35] is a widely used unsupervised classification technique which partitions the input space into K regions via some similarity measures. The partition is done in such a way that patterns within a group are more similar to each other than those belonging to different groups. There are two main kinds of clustering algorithms, i.e. partitioning and hierarchical, available in the literature. Partitioning algorithms, like K-means algorithm [4,38], attempt to create k partitions in the situation that k is fixed and known in advance. K-means is

* Corresponding author. Tel.: +86 0510 85326695; fax: +86 0510 85326691.
E-mail address: songwei@jiangnan.edu.cn (W. Song).

http://dx.doi.org/10.1016/j.ins.2014.03.024
0020-0255/

an iterative hill-climbing algorithm which suffers from the restriction of suboptimum depending on the choice of initial clustering distribution. Some, like K-means algorithm, optimize of the distance criterion either by minimizing the within cluster spread, or by maximizing the inter-cluster separation. Hierarchical algorithms, like OPTICS [1], compute a representation of the possible hierarchical cluster structure from which clusters at various solutions can be extracted. Other techniques, like graph theoretical approach [19], branch and bound procedure [18], maximum likelihood estimate technique algorithm [44], etc., perform clustering based on other criteria and have proved to be useful in some specific applications of computational intelligence. In recent decades, several types of biologically inspired clustering algorithms have been proposed for clustering. Ant clustering algorithm [20,47] projects the original data into bidimensional output grids and positions that are similar to each other in their original attributes. By doing so, the algorithm is capable of grouping items together that are similar to each other. Genetic algorithm (GA) [5,34,36,40] is a randomized search and optimization technique which can be used in complex and large landscapes and provide near-optimal solutions for its optimization problem. However, when these algorithms are applied in text clustering, most of them solely use vector space model (VSM) to represent text, that is to say, each unique word in vocabulary represents one dimension in vector space [29]. The representation by pack of words adopted in these clustering algorithms is often unsatisfactory, because VSM makes matches simply via keywords. However, the same concept can be signified by different words. Thus, the relationship between some important words which do not co-occur literally are ignored in VSM. Moreover, the inherent high dimensional space with a large number of features leads to a high cost of computational time for clustering problems. In this paper, a fuzzy control genetic algorithm is proposed, in conjunction with a hybrid semantic similarity measure, for document clustering. We use the broad-coverage taxonomy and hierarchical structure of WordNet [15,25] as thesaurus-based ontology to detect semantic relationships between documents. Moreover, a new semantic space model (SSM) is proposed and demonstrated as a corpus-based method which can appropriately depict the associative semantic relationships. Moreover, we greatly reduce the number dimensions for document representation. Considering the mutual influence between the selection pressure and the diversity of population, we recommend a fuzzy control GA which takes advantage of dynamic probabilities of crossover and mutation to realize the goals of GA, i.e. maintaining the diversity as well as sustaining the convergence capacity. As is known, the parameters of GA critically affect its success and are rather difficult to define. In our method, they are adjusted adaptively by GA itself.

The following parts of this paper are organized as below: Section 2 describes how to calculate semantic similarity via thesaurus-based method. In Section 3 a semantic space model (SSM) is proposed and demonstrated, in conjunction with the thesaurus-based measure as a hybrid strategy for semantic similarity calculation. The details of fuzzy control genetic algorithm based on the semantic similarity measure for document clustering are given in Section 4. Experiment results and analysis are given in Section 5. Conclusions are given in Section 6.

2. Semantic similarity based on ontology

Semantic similarity is a generic method adopted in the fields of information retrieval (IR) and artificial intelligence (AI). Semantic similarity between two words is often represented by the similarity between the concepts associated with the two words. Thus, it can overcome the limitations of the simple cosine measure based on the bag-of-words representation of documents in VSM. Several semantic similarity methods have been developed in the specific applications of computational intelligence [21,31]. In general, these measures can be categorized into two kinds: thesaurus-based methods and corpus-based methods.

2.1. WordNet: A widely used thesaurus

WordNet is a widely used semantic dictionary developed under the direction of Miller [25]. It organizes words, i.e. nouns, verbs, adjectives and adverbs, by synonym sets, named synsets, each representing a distinct concept. Synsets are interlinked via conceptual-semantic and lexical relations [22]. The version utilized in this paper is WordNet 2.0, which consists of 109,377 synsets and 144,684 words.

2.2. Thesaurus-based semantic similarity calculation

In this paper we adopt the thesaurus-based method given by Li et al. [21]. In their study two factors for calculating the similarity between two concepts are taken into account: (1) The shortest path length between the two concepts and (2) the depth of the subsumer in the hierarchical structure. That is to say, given two concepts c_1 and c_2, the semantic similarity is defined as:

$$sim(c_1, c_2) = f_1(l) \cdot f_2(h) \tag{1}$$

where l is the shortest path length connecting concept c_1 and c_2. h is the depth of subsumer in the hierarchical structure. Suppose the respective effects of l and h on the similarity are independent from each other, therefore, the similarity function is comprised of two independent functions of f_1 and f_2.

Suppose a word has multiple meanings, therefore, different paths may exist and the minimum length of the path connecting the two concepts is a direct method to calculate the similarity. It is intuitive that the similarity between two concepts